METHODS FOR THE ANALYSIS OF DATA FROM MULTIVARIATE DIRECTED GRAPHS*

by

Stephen E. Fienberg
Stanley Wasserman

Technical Report #351

Department of Applied Statistics
School of Statistics
University of Minnesota
St. Paul, MN 55108

November, 1979

*This research was partially supported by NSF Grant #SOC78-26075. This paper was presented at a conference on "Recent Developments in Statistical Methods and Applications", December 1979, sponsored by the Institute of Mathematics, Academia Sinica, Taipai, Taiwan, Republic of China. We thank David Hinkley for helpful conversations, and Michael Meyer and Rick Picard for computational assistance.

Abstract

A multivariate directed graph consists of a set of g nodes, and a family of directed arcs (one for each relation) connecting pairs of nodes. Such multivariate directed graphs provide natural representations for social networks. In this paper we describe a class of stochastic loglinear models for multivariate directed graphs, demonstrate how they can be fit to sociometric data, and explain the links between these models and standard loglinear models for multidimensional contingency tables. We apply the models to data on the relationships among 73 organizations in a Midwest community.

1. Introduction

Social networks, linking individuals, groups, or organizations by means of various social or behavioral relationships, have increasingly been used as a paradigm for studying social structure (see e.g., the collection of articles in Leinhardt, 1977). Until recently these networks typically were modelled in a nonstochastic fashion by means of univariate and multivariate directed graphs. In this paper we describe a class of stochastic models for multivariate directed graphs, demonstrate how these models can be fit to sociometric data, and explain the links between these models and standard loglinear models for multidimensional contingency tables (e.g., see Bishop, Fienberg, and Holland, 1975, and Fienberg, 1977).

In the next section, we give some basic notation for directed graphs, and introduce the first part of the assumed stochastic structure. Section 3 contains a description of a study reported on in Galaskiewicz and Marsden (1978) and Galaskiewicz (1979) on the relationships among organizations in a small Midwest American community. Then, in Section 4, we describe a class of loglinear models suitable for the analysis of the Galaskiewicz-Marsden data. In Section 5, we describe how to fit these models, and we link them to variants of the standard loglinear models for multidimensional contingency tables, which we then fit to the data in Section 6.

Finally, in Section 7, we discuss extensions of the models of Section 4, and describe how they relate to a class of models for univariate directed graphs described in Holland and Leinhardt (1979), and in Fienberg and Wasserman (1979).

2. Directed Graphs

A directed graph, or digraph, consists of a set of g nodes, and a set of directed arcs connecting pairs of nodes. Digraphs are natural mathematical representations of social networks, where the nodes represent individuals or organizations and the arcs represent directed relationships, such as friendship. Figure 1 is an illustration of a digraph with 6 nodes ($g = 6$) and 12 directed arcs. This figure is one of a univariate digraph with the arcs describing a single type of relationship, referred to as a "generator", for which the maximum number of arcs is $g(g - 1) = 6 \times 5 = 30$.

A digraph, D_g, with g nodes can thus be summarized by means of a $g \times g$ sociomatrix or adjacency matrix $\underset{\sim}{x}$, with

$$x_{ij} = \begin{cases} 1 & \text{if i relates to j} \\ 0 & \text{otherwise.} \end{cases}$$

By convention, the g diagonal terms x_{ii} are set equal to 0. The digraph for Figure 1 is given by the matrix:

$$\underset{\sim}{x} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \left[\begin{array}{cccccc} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array} .$$

The matrix $\underset{\sim}{x}$ can be thought of as the realization of a matrix of random variables, $\underset{\sim}{X}$, where we assume that the $\binom{g}{2}$ pairs or dyads,

$$D_{ij} = (X_{ij}, X_{ji}), \quad i > j,$$

are independent bivariate random variables, with $2^2 = 4$ possible realizations:

$$D_{ij} = \begin{cases} (1,1) : \text{Mutual} \\ (1,0) \text{ or } (0,1) : \text{Asymmetric} \\ (0,0) : \text{Null.} \end{cases}$$

A multivariate directed graph, or multigraph, consists of a set of g nodes, a family of n binary relations or generators, $R_1, R_2, \ldots, R_n$, and n sets of directed arcs connecting pairs of nodes, where the rth set corresponds to R_r. The multigraph is described by the collection of random variable sociomatrices $\mathbf{X} = \{\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n\}$, and we assume that the $\binom{g}{2}$ dyads,

$$\underset{\sim}{D}_{ij} = \begin{pmatrix} X_{ij1}, & X_{ji1} \\ X_{ij2}, & X_{ji2} \\ \vdots & \vdots \\ X_{ijn}, & X_{jin} \end{pmatrix}, \quad i > j,$$

are independent 2n-variate random variables with 2^n possible realizations. When n = 3, the $\underset{\sim}{D}_{ij}$ are sex-variate random variables with $2^6 = 64$ possible realizations. In the next section, we note that only 36 of these 64 realizations are distinguishable.

For both digraphs and multigraphs the assumption that the dyads are independent random variables is a crucial one, and is not subject to examination within the framework developed in this paper. We comment further on this assumption in Fienberg and Wasserman (1979).

3. The Galaskiewicz-Marsden Data

Galaskiewicz and Marsden (1978) report on a study of the social structure linking the 109 identifiable organizations in a small Midwest community of 32,000 persons. These organizations included all manufacturing firms having more than 20 employees, banks, savings and loans, law firms, business associations, service clubs, labor unions, city offices and departments, political organizations, mass media organizations, health institutions, public welfare institutions, educational institutions, and churches.

Agents from 73 of the 109 organizations were interviewed and the following list of six questions was asked:

A. Which organizations on this list does (your organization) rely upon for information regarding community affairs (or other matters that might affect it)?
B. And to which organizations on this list would (your organization) be likely to pass on important information concerning community affairs (or other matters that might affect them)?
C. Now to which organizations on this list does (your organization) give substantial funds as payment for services rendered or goods received, loans, or donations?
D. And from which organizations on this list does (your organization) get substantial funds as payment for services rendered or goods provided, loans or donations?
E. Which organizations on this list does (your organization) feel a special duty to stand behind in time of trouble: that is, to which organizations would (your organization) give support?
F. Finally, which organizations on this list would be likely to come to (your organization's) support in time of trouble?

From the responses to these questions, Galaskiewicz and Marsden constructed three 73 × 73 adjacency matrices, one each for information, money, and support.

They took a flow in a given direction between a pair of organizations to be present if the agent for either organization responded positively to the related question. Then they summarized the dyadic information in the form of a 2^6 cross-classification, reproduced here as Table 1.

For Table 1, each dyad was counted twice, once from the perspective of each member of the pair. Thus the total in the table, 5256, is twice the number of dyads, $\binom{73}{2} = 2628$. This double counting leads to the duplication of all 28 counts corresponding to those dyads, D_{ij}, with at least one asymmetric row (i.e., (1,0) or (0,1) for at least one of information, money or support), and the doubling of the 8 counts corresponding to dyads whose rows are either all mutual (i.e., (1,1)), or all null (i.e., (0,0)) or a mixture of mutuals and nulls.

While the duplication and doubling of counts in Table 1 appears both wasteful and possibly incorrect, we will find Table 1 of special use for the compution of maximum likelihood estimates corresponding to the models developed in the next section. Thus we can view the double counting as a computational device, similar to the one used by Bishop, Fienberg, and Holland (1975, Chapter 8) for the analysis of square contingency tables using the model of quasi-symmetry.

Following Galaskiewicz and Marsden, we propose to study the organizational network summarized in Table 1 in terms of "patterns" for the relationships between a randomly selected pair of organizations, say A and B. They suggest the six basic patterns of "association", depicted in Figure 2. In the next section we incorporate parameters in our loglinear models corresponding to these patterns, thus providing explicit interpretations for the notion of "association".

Note that the data summary in Table 1 aggregates across dyads, treating all 2628 dyads alike, and thereby ignoring both (a) effects due to the

specific organizations involved in each dyad, and (b) information that we have available regarding the organizations, such as whether they are banks, business associations, churches, labor unions, etc. Holland and Leinhardt (1979) refer to (b) as nodal attributes. Other analyses, that we have carried out on the basic adjacency matrices from which Table 1 was constructed (Fienberg and Wasserman, 1979), suggest that ignoring this information is a gross oversimplification. Nonetheless, we proceed to develop models for this simplified structure, and only return to a discussion of the more complex problem involving this extra information in the final section of the paper.

4. Loglinear Models for Multigraphs

In order to model the multivariate directed graph data of Galaskiewicz and Marsden, we need to develop a representation for the summary counts adding across dyads, without the duplications in Table 1. We give such a representation here in a form resembling a three-dimensional $4 \times 4 \times 4$ cross-classification, where the three variables correspond to the three generators, information, money, and support.

When the dyadic structure for only a single generator is asymmetric, the "direction" of the corresponding arc does not matter. We use a single subscript, A, for the corresponding generator in such situations. When the dyadic relationship for two or more generators is asymmetric, we need to distinguish between situations where the arcs for a pair of generators go in the same or different directions. Thus, for these situations, we use two different subscripts, A and $\overline{A}$, with identical subscripts for those generators whose asymmetric directed arcs go in the same direction, and we arbitrarily assign the subscript A to the lowest numbered asymmetric generator. (Note that interchanging the subscripts A and $\overline{A}$ yields the same dyadic structural relationship.)

We denote the counts in this table by Z_{abc}, where a, b, c = M, A, $\overline{A}$, N (for Mutual, Asymmetric, $\overline{A}$symmetric, and Null), and where the convention for the use of the subscripts A and $\overline{A}$ is as described above. In Table 2, we give the representation of these counts in the form of a $4 \times 4 \times 4$ cross-classification, where there is a collapsing for those cells corresponding to dyadic structures with asymmetrics for generators 2 and/or 3, but not generator 1. There are 36 cells in this array. Table 2 also contains the unduplicated set of counts corresponding to the data in Table 1.

The sampling structure here corresponds to a multinomial model where the $N = \binom{73}{2} = 2628$ dyads are independently assigned to the 36 cells of Table 2, according to a set of underlying cell probabilities $\{p_{abc}\}$ where $\sum_{\text{all cells}} p_{abc} = 1$. Let

$$\xi_{abc} = \log p_{abc}. \tag{4.1}$$

We wish to develop a class of increasingly complex loglinear models for the $\{\xi_{abc}\}$, based on sets of parameters corresponding to the association patterns depicted in Figure 2:

(i) θ - - - grand mean (or normalization constant),

e.g.: $\xi_{abc} = \theta$, for all a, b, c.

(ii) $\theta_1, \theta_2, \theta_3$, - - - choice parameters corresponding to the presence of directed arcs for each generator,

e.g.: $\xi_{MAN} = \theta + 2\theta_1 + \theta_2$,

$\xi_{MAA} = \theta + 2\theta_1 + \theta_2 + \theta_3$,

$\xi_{MA\bar{A}} = \theta + 2\theta_1 + \theta_2 + \theta_3$.

(iii) $\rho_{11}, \rho_{22}, \rho_{33}$ - - - symmetry effects corresponding to Figure 2a,

e.g.: $\xi_{MA\bar{A}} = \theta + 2\theta_1 + \theta_2 + \theta_3 + \rho_{11}$

(iv) $\rho_{12}, \rho_{13}, \rho_{23}$ - - - exchange effects involving pairs of generators and corresponding to Figure 2b,

e.g.: $\xi_{MAA} = \theta + 2\theta_1 + \theta_2 + \theta_3 + \rho_{11} + \rho_{12} + \rho_{13}$,

$\xi_{MA\bar{A}} = \theta + 2\theta_1 + \theta_2 + \theta_3 + \rho_{11} + \rho_{12} + \rho_{13} + \rho_{23}$.

(v) $\gamma_{12}, \gamma_{13}, \gamma_{23}$ - - - multiplexity effects corresponding to Figure 2c,

e.g.: $\xi_{MAA} = \theta + 2\theta_1 + \theta_2 + \theta_3 + \rho_{11} + \rho_{12} + \rho_{13} + \gamma_{12} + \gamma_{13} + \gamma_{23}$,

$$\xi_{MA\bar{A}} = \theta + 2\theta_1 + \theta_2 + \theta_3 + \rho_{11} + \rho_{12} + \rho_{13} + \rho_{23} + \gamma_{12} + \gamma_{13}.$$

(vi) $(\rho\gamma)_{112}$, $(\rho\gamma)_{113}$, $(\rho\gamma)_{221}$, $(\rho\gamma)_{223}$, $(\rho\gamma)_{331}$, $(\rho\gamma)_{332}$ - - - conditional asymmetry effects (involving exchange conditional on symmetry or symmetry conditional on either exchange or multiplexity), corresponding to Figure 2d,

e.g.: $$\xi_{MA\bar{A}} = \theta + 2\theta_1 + \theta_2 + \theta_3 + \rho_{11} + \rho_{12} + \rho_{13} + \rho_{23} + \gamma_{12} + \gamma_{13} + (\rho\gamma)_{112} + (\rho\gamma)_{113}.$$

(vii) $(\rho\gamma)_{1122}$, $(\rho\gamma)_{1133}$, $(\rho\gamma)_{2233}$ - - - multiplex symmetry effects corresponding to Figure 2e,

e.g.: $$\xi_{MM\bar{A}} = \theta + 2\theta_1 + 2\theta_2 + \theta_3 + \rho_{11} + \rho_{22} + 2\rho_{12} + \rho_{13} + \rho_{23} + 2\gamma_{12} + \gamma_{13} + \gamma_{23} + 2(\rho\gamma)_{112} + 2(\rho\gamma)_{221} + (\rho\gamma)_{113} + (\rho\gamma)_{223} + (\rho\gamma)_{1122}.$$

We note that these models are structured to <u>be hierarchical</u>, in that, if a term is set equal to zero, all of its "higher-order relatives" must also be set equal to zero, e.g.:

$$\theta_1 = 0 \text{ implies } \rho_{11} = \rho_{13} = \gamma_{12} = \gamma_{13} = (\rho\gamma)_{112} = (\rho\gamma)_{113} = (\rho\gamma)_{221} = (\rho\gamma)_{331} = (\rho\gamma)_{1122} = (\rho\gamma)_{1133} = 0; \quad (4.2)$$

and

$$\rho_{12} = 0 \text{ implies } (\rho\gamma)_{112} = (\rho\gamma)_{221} = (\rho\gamma)_{1122} = 0. \quad (4.3)$$

Although the ξ's have only three subscripts, the $\{\rho_{ij}\}$, $\{\rho_{ij}|i>j\}$, and $\{\gamma_{ij}|i>j\}$ can all be thought of as "two factor" effects, the $\{(\rho\gamma)_{iij}|i\neq j\}$ as "three-factor" effects, and the $\{(\rho\gamma)_{iijj}|i\neq j\}$ as "four-factor" effects,

with respect to the 6 factors or dimensions in Table 1.

It is easy to formulate additional parameters that could be used in models for the $\{\xi_{abc}\}$, but those described above are sufficient to illustrate our approach, and to provide a good fit to the data in Table 2. We note that, except for the constraint, (4.1), stating that p_{abc}'s add to 1, there are no constraints on the loglinear model parameters in (i) through (vii). Each parameter corresponds to the presence of a particular combination of directed arcs (as depicted in Figure 2), and thus the notation is closer to that of Nelder and Wedderburn (1972) as implemented in the GLIM package of computer programs, than it is to the u-term notation of Bishop, Fienberg, and Holland (1975). Such a choice of notation has no effect on the computation of estimated expected cell values,

$$\hat{m}_{abc} = N\hat{p}_{abc},$$

as described in the following section.

5. Computing Estimated Expected Values

The models given in the previous section are loglinear models for a multinomial sampling problem, so that general results for loglinear models given by Haberman (1974), and outlined in Fienberg (1977, Appendix II) are directly applicable.

The two key results from the general theory are as follows:

A. For a loglinear model corresponding to the subspace $\mathcal{M}$, the minimal sufficient statistics (MSS's) are given by the projection of the vector of counts, $\underset{\sim}{Z}$, onto $\mathcal{M}$, denoted by $P_{\mathcal{M}}\underset{\sim}{Z}$. These MSS's take the form of linear combinations of the components of $\underset{\sim}{Z}$.

B. The maximum likelihood estimates (MLE's) for the vector of expected values, $\underset{\sim}{m}$, if they exist, are found by setting the MSS's equal to their expected values, i.e.,

$$P_{\mathcal{M}}\,\underset{\sim}{\hat{m}} = P_{\mathcal{M}}\,\underset{\sim}{Z}. \tag{5.1}$$

The MSS's for the parameters in each of the hierarchical loglinear models introduced in Section 4, are linear combinations of the observed counts $\{Z_{abc}\}$ where the multiplier for each cell is either 0, 1, or 2. (The 2's appear in conjunction with those cells where the model for ξ_{abc} has a factor of 2 multiplying the relevant parameter.) In Table 3 we give examples of the MSS's for each of the 6 types of parameters.

For some of the models of Section 4, the MSS's can be written simply as marginal totals of Table 2. For example, for the model with parameters

$$\{\theta,\theta_1,\theta_2,\theta_3,\rho_{11},\rho_{22},\rho_{33}\}, \tag{5.2}$$

the MSS's consist of

$$Z_{+++},$$

$$2Z_{M++} + Z_{A++},\ 2Z_{+M+} + Z_{+A+},\ 2Z_{++M} + Z_{++A},$$

$$Z_{M++},\qquad Z_{+M+},\qquad Z_{++M},$$

or equivalently, by taking appropriate differences,

$$\{Z_{a++};\ Z_{+b+};\ Z_{++c} \mid a,b,c = M,A,N\}, \tag{5.3}$$

where, for example, Z_{+A+} is the sum of all cell counts for which the 2nd generator is asymmetric. (This notation reintroduces the symmetry absent in Table 2.) Thus the MLE's for model (5.2) are the solution of the equations

$$\hat{m}_{a++} = Z_{a++}, \qquad a = M,A,N\ , \tag{5.4}$$

$$\hat{m}_{+b+} = Z_{+b+}, \qquad b = M,A,N\ , \tag{5.5}$$

$$\hat{m}_{++c} = Z_{++c}, \qquad c = M,A,N\ . \tag{5.6}$$

The likelihood equations in (5.4), (5.5), and (5.6) are written in a symmetric form and thus contain some redundancies.

On the other hand, the model with parameters

$$\{\theta,\theta_1,\theta_2,\theta_3,\rho_{11},\rho_{22},\rho_{33},\rho_{12},\rho_{13},\rho_{23}\} \tag{5.7}$$

has MSS's which do not reduce to simple sums. In addition to the sums in (5.3), they include the following linear combinations.

$$2Z_{MM+} + Z_{MA+} + Z_{AM+} + Z_{A\bar{A}+},\quad 2Z_{MN+} + Z_{MA+} + Z_{AN+} + Z_{AA+} \tag{5.8}$$

plus two other such pairs corresponding to ρ_{23} and ρ_{13}.

Note that, for both model (5.2) and model (5.7) as well as all other hierarchical loglinear models from Section 4, the MSS's are equal or

equivalent to marginal totals for the 2^6 table of duplicated and doubled counts corresponding to Table 1, with entries $\{Y_{ii'jj'kk'}\}$. In this Y-table each subscript is 1 if the relation is present and 0 if it is not. Indeed, we can think of Table 1 as a "pseudo-table" constructed to have marginal totals equal to sets of these MSS's. Since specifying marginal totals is a common way to specify loglinear models for multidimensional contingency tables, we can associate with the loglinear model for the Z-table a unique loglinear model for the Y-table whose MSS's "match" those of the model for the Z-table. We need to be very careful in specifying the interaction structure of the model, however, as we note below.

It is then natural to ask:

> Are the estimated expected values for the two loglinear models (one for the Z-table and the associated one for the Y-table) the same, once we take into account the doubling and duplication?

The answer to this question is yes. The proof of this equivalence involves two steps. First, we need to show that the likelihood equations for the Z-table are the same as for the Y-table. Second, we need to show that the estimated expected values from the Y-table satisfy the loglinear model for the Z-table. We illustrate by example.

Consider the loglinear model for the Z-table with parameters given by (5.2) and likelihood equations by (5.4), (5.5), and (5.6). First we note that

$$\begin{aligned} Z_{M++} &= \frac{1}{2} Y_{11++++}, \\ Z_{A++} &= Y_{10++++} = Y_{01++++}, \\ Z_{N++} &= \frac{1}{2} Y_{00++++}, \end{aligned} \tag{5.9}$$

with similar equivalences for Z_{+b+} and Z_{++c} for b,c = M,A,N. Thus the MSS's correspond to the marginal totals

$$\{Y_{ii'++++}\},\ \{Y_{++jj'++}\},\ \text{and}\ \{Y_{++++kk'}\}. \tag{5.10}$$

The loglinear model for the Y-table for which these are MSS's and for which the likelihood equations reduce to (5.4), (5.5), and (5.6) is given by

$$\begin{aligned}\log m_{ii'jj'kk'} = {} & \lambda + \lambda_1\delta_i + \lambda_{1'}\delta_{i'} + \lambda_2\delta_j + \lambda_{2'}\delta_{j'} \\ & + \lambda_3\delta_k + \lambda_{3'}\delta_{k'} + \lambda_{11}\delta_i\delta_{i'} \\ & + \lambda_{22}\delta_j\delta_{j'} + \lambda_{33}\delta_k\delta_{k'} + \delta^*_{ii'jj'kk'}\log 2\end{aligned} \tag{5.11}$$

where each δ-term equals 1 if the subscript takes the value 1, and is zero otherwise, and

$$\delta^*_{ii'jj'kk'} = \begin{cases} 1 & \text{if } i = i',\ j = j',\ \text{and } k = k' \\ 0 & \text{otherwise.} \end{cases} \tag{5.12}$$

The final term in (5.11) involving δ^* is required to account for the 8 doubled cell counts in Table 1; since this term does not involve any parameters the MSS's are the same as those for the standard loglinear model for the 2^6 table, i.e., the two-way marginal totals in (5.10). Were it not for the doubling of the 8 cell counts, we would not have to include the δ^* term, and then (5.11) would be identical to a standard loglinear model for the 2^6 table. The inclusion of this final term induces a 6-factor interaction term into the model, and thus we can not use the standard iterative procedures.

Finally we note that, because of the symmetries in the marginal totals of the Y-table, i.e.,

$$Y_{10++++} = Y_{01++++},$$

$$Y_{++10++} = Y_{++01++},$$

$$Y_{++++10} = Y_{++++01},$$

the estimated expected values are such that

$$\hat{\lambda}_1 = \hat{\lambda}_{1'}, \quad \hat{\lambda}_2 = \hat{\lambda}_{2'}, \quad \text{and} \quad \hat{\lambda}_3 = \hat{\lambda}_{3'}.$$

It then follows that we have a 1-1 correspondence between the estimated parameters associated with (5.11) and those of the estimated parameters in the loglinear model for the Z-table:

$$\hat{\lambda} = \hat{\theta} + \log 2N,$$

$$\hat{\theta}_1 = \hat{\lambda}_1 = \hat{\lambda}_{1'}, \ \hat{\theta}_2 = \hat{\lambda}_2 = \hat{\lambda}_{2'}, \ \hat{\theta}_3 = \hat{\lambda}_3 = \hat{\lambda}_{3'},$$

$$\hat{\rho}_{11} = \hat{\lambda}_{11}, \quad \hat{\rho}_{22} = \hat{\lambda}_{22}, \quad \hat{\rho}_{33} = \hat{\lambda}_{33},$$

where $N = \binom{g}{2} = \binom{73}{2}$, the total number of dyads. The term, log 2N, on the right-hand side in the first of these equalities is required because we have written (5.11) as a loglinear model for the expected values of the Y-table, whereas the models for the $\{\xi_{abc}\}$ in Section 4 are for the cell probabilities for the Z-table. Thus we can fit (5.11) to the Y-table and produce the appropriate estimated expected values and estimated parameters for the model associated with the Z-table.

While the equivalence of expected values noted above is for a specific model from Section 4, similar equivalences hold for all other models considered there. The correspondence between models for the Z-table and models for the Y-table is summarized in Table 4, where we use Fienberg's (1977) notation, [22'], [12], and [112'], etc., to refer to the marginal totals of the Y-table that are the MSS's for the loglinear model. This notation is slightly misleading, however, since all of the models to be fit also include the 6-factor interaction term corresponding to the final term in (5.11).

The method we use to solve the likelihood equations is a variant on the standard iterative proportional scaling algorithm applied to the Y-table, where we take as our initial values

$$\hat{m}^{(0)}_{ii'jj'kk'} = 1 + \delta_{ii'jj'kk'}. \tag{5.13}$$

Thus the initial values are 1 in each of the duplicated cells, and are 2 in each of the doubled cells. Then we proceed to adjust for each of the fitted margins in turn in the usual fashion for multidimensional tables (e.g., see Bishop, Fienberg, and Holland, 1975; or Fienberg, 1977). This iterative procedure will preserve the 6-factor interaction introduced by the initial values, (5.13). Simply applying the standard iterative scaling algorithm with initial values equal to 1 to this table yields incorrect answers as model (5.11) is not satisfied. We note that many iterative scaling programs allow the specification of initial values such as (5.13).

If a likelihood ratio statistic, G^2, is computed for the fit of the model to the Y-table as in a standard contingency table program, the value so computed must be divided by a factor of 2 since all of the counts are either duplicated or doubled. Alternatively G^2 can be computed directly as:

$$G^2 = 2 \sum_{\text{all cells}} Z_{abc} \log \frac{Z_{abc}}{\hat{m}_{abc}} . \qquad (5.14)$$

These G^2 values can then be referred to reference chi-square distributions on the appropriate degrees of freedom. Note that both the degrees of freedom and the estimated parameter values from the standard contingency table program output are incorrect, since they do not take into account the double-counting. The degrees of freedom are given by

$$df = (\#\text{ cells in Z-table}) - (\#\text{ parameters fitted}), \qquad (5.15)$$

and are listed as the final column in Table 4. The estimated θ, ρ, γ, and $(\rho\gamma)$ parameters can be computed by taking appropriate contrasts of the $\{\log \hat{m}_{abc}\}$.

Why have we gone to so much trouble to note the equivalence of estimated expected values computed for the Z-table and the Y-table? The answer is because of the form of the MSS's for the Z-table, i.e., because for several models the MSS involve weights for some cell counts that are double those for others. Thus, to compute estimated expected cell values for these models directly from the Z-table, we need to use the generalized iterative scaling (GIS) algorithm of Darroch and Ratcliff (1972), as described in Fienberg (1977). This algorithm involves steps in which complicated multiplicative adjustments are made involving powers of $\frac{1}{2}$, and its convergence is considerably slower than the convergence of our variant of the iterative scaling algorithm applied to the Y-table, described above. A rather intricate argument, not reproduced here, can be used to demonstrate the equivalence of the GIS algorithm for the Z-table, and the variant of iterative scaling algorithm for the Y-table.

6. Analysis of the Galaskiewicz-Marsden Data

We now turn to a discussion of our analysis of the Galaskiewicz-Marsden data in Table 1. In Table 5, we summarized the fit of 6 of the loglinear models listed in Table 4. The only model not included is the null model, (i).

The only model that provides an adequate fit to these data is the full multiplex symmetry model, (vii), and its G^2 value is barely greater than the $\chi^2_{14}(.05)$ value (23.7). We list the estimated parameter values for model (vi) in Table 6.

The G^2 values in Table 5 differ from those values reported in Galaskiewicz and Marsden (1978). They used incorrect initial values for the iterative scaling procedure, and thus did not compute the estimated expected values which are MLE's for the models of Section 4. The estimated parameter values reported in their paper are also incorrect.

We conclude by noting that more complex models can be devised which do fit the data well. Indeed the model with fitted margins

[11'22'3] [11'22'3'] [122'33'] [1'22'33'] [11'33']

has a value of $G^2 = 2.4$.

7. Extensions

A natural extension to the models considered in this paper involves adding parameters corresponding to each of the "nodes" or organizations. In particular, as we noted in Section 3, the data in Table 1, and consequently our analyses of these counts, aggregate across dyads, and thereby ignore effects due to the specific organizations involved in each dyad. Holland and Leinhardt (1979) develop a model for single generators with individual parameters, which they label p_1. We (Fienberg and Wasserman, 1979) have considered p_1 in situations where we have additional data on the nodes. For example, with organizations as nodes, we might know whether each organization was local or "extra-local", or whether it was private or public. We briefly describe p_1 and its extensions utilizing such nodal data in this section, and suggest how one might build a multivariate version of p_1 to deal with a more complete version of the Galaskiewicz-Marsden data.

Suppose we have a single sociomatrix, $\underset{\sim}{X}$, and consider all $N = \binom{g}{2}$ dyads. The Holland-Leinhardt p_1 density function postulates that:

$$\begin{aligned}
&\ln P\{X_{ij} = 0, X_{ji} = 0\} = \lambda_{ij}, \\
&\ln P\{X_{ij} = 1, X_{ji} = 0\} = \lambda_{ij} + \alpha_i + \beta_j + \theta, \\
&\ln P\{X_{ij} = 0, X_{ji} = 1\} = \lambda_{ij} + \alpha_j + \beta_i + \theta, \\
&\ln P\{X_{ij} = 1, X_{ji} = 1\} = \lambda_{ij} + \alpha_i + \alpha_j + \beta_i + \beta_j + 2\theta + \rho,
\end{aligned} \tag{7.1}$$

subject to the constraints that these four joint probabilities sum to 1 for every dyad, and that

$$\alpha_+ = \beta_+ = 0. \tag{7.2}$$

The $\{\alpha_i\}$ and $\{\beta_j\}$ parameters measure respectively, the "expansiveness of the nodes, and the "popularity" of the nodes, while ρ is a measure of reciprocity. These parameters reflect the individual effects due to the specific nodes in each dyad. The $\{\lambda_{ij}\}$ are "normalizing constants" and are unrelated to the parameters in the models of Sections 4 and 5. The parameters θ and ρ play the same role as θ_j and ρ_{jj}, $j = 1,2,3$ in the three generator model of Section 4.

The MSS's for the parameters of p_1 are:

ρ	$\sum_{i>j} X_{ij}X_{ji}$	Number of mutuals,
α_i	X_{i+}	Outdegree of node i $i = 1,2,\ldots,g$,
β_j	X_{+j}	Indegree of node j $j = 1,2,\ldots,g$,
θ	X_{++}	Number of choices.

Calculation of estimated expected values for p_1 and related models is considered in Fienberg and Wasserman (1979) and involves construction of a pseudo-table of counts not unlike the approach utilized in this paper.

Next, suppose we classify the nodes into subgroups, so that all nodes in a given subgroup have identical scores on a set of nodal variables. For the nodes in each subgroup, we then might choose to equate the $\{\alpha_i\}$ and $\{\beta_j\}$ parameters, i.e., we take these nodes to have a common α and a common β which measure the expansiveness and popularity of the subgroup as a whole. This variant on p_1 is also discussed in Fienberg and Wasserman (1979).

The foregoing discussion leads naturally to a meshing of a multivariate version of p_1 with the models of Section 4. Since ρ_{rr} and θ_r (for $r = 1,2,3$) correspond to ρ and θ in p_1, and thus are already present in our models, we can think of adding individual parameters to our models for each generator, i.e., α_{ri} and β_{ri} for $r = 1, 2,\ldots,n$ (the number of generators), and $i = 1,2,\ldots,g$. Since $g = 73$ in our example, this is an overwhelming number of new parameters, and we can equate the α's and β's within groups as we suggested above for a single generator. We plan to discuss these extensions more fully, and apply them to the Galaskiewicz-Marsden data in the near future.

8. References

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland (1975), Discrete Multivariate Analysis. Cambridge, MA: The MIT Press.

Darroch, J.N. and D. Ratcliff (1972), "Generalized iterative scaling of log-linear models," The Annals of Mathematical Statistics, 43:1470-1480.

Fienberg, S.E. (1977), The Analysis of Cross-Classified Categorical Data. Cambridge, MA: The MIT Press.

Fienberg, S.E. and S. Wasserman (1979), "Categorical data analysis of single sociometric relations," Technical Report #362, School of Statistics, University of Minnesota.

Galaksiewicz, J. and P.V. Marsden (1978), "Interorganizational resource networks: Formal patterns of overlap," Social Science Research, 7: 89-107.

Galaskiewicz, J. (1979), Exchange Networks and Community Politics. Beverly Hills, CA: Sage Publications.

Haberman, S. (1974), The Analysis of Frequency Data. Chicago: The University of Chicago Press.

Holland, P.W. and S. Leinhardt (1979), "An exponential family of probability densities for directed graphs," Unpublished manuscript.

Leinhardt, S. (1977), Social Networks: A Developing Paradigm. New York: Academic Press.

Nelder, J.A. and R.W. Wedderburn (1972), "Generalized linear models," Journal of the Royal Statistical Society, Series A, 135:370-384.

Table 1

Observed Distribution of Interorganizational Transactions Involving Three Resources and 73 Organizations[a] (Source: Galaskiewicz and Marsden (1978))

1. Information out		-								+							
1'. Information in		-				+				-				+			
2. Money out		-		+		-		+		-		+		-		+	
2'. Money in		-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+
3. Support out	3'. Support in																
-	-	3020	89	89	24	145	16	17	8	145	17	16	8	332	47	47	16
	+	115	17	11	3	21	9	4	4	31	18	2	1	77	37	16	25
+	-	115	11	17	3	31	2	18	1	21	4	9	4	77	16	37	25
	+	110	13	13	4	19	4	7	0	19	7	4	0	102	52	52	32

Table Total = 5256

[a] "+" indicates that a directed flow is present, "-" indicates that a directed flow is absent.

Table 2

STRUCTURE FOR ACTUAL TABLE OF COUNTS, ALONG WITH UNDUPLICATED COUNTS CORRESPONDING TO TABLE 1.

				GENERATOR 3			
				M	A		N
GENERATOR 1	M	GENERATOR 2	M	Z_{MMM} 16	Z_{MMA} 25		Z_{MMN} 8
			A	Z_{MAM} 52	Z_{MAA} 37	$Z_{MA\bar{A}}$ 16	Z_{MAN} 47
			N	Z_{MNM} 51	Z_{MNA} 77		Z_{MNN} 166
	A	GENERATOR 2	M	Z_{AMM} 0	Z_{AMA} 4	$Z_{AM\bar{A}}$ 1	Z_{AMN} 8
			A	Z_{AAM} 4	Z_{AAA} 9	$Z_{AA\bar{A}}$ 2	Z_{AAN} 16
				$Z_{\bar{A}AM}$ 7	$Z_{\bar{A}AA}$ 4	$Z_{\bar{A}A\bar{A}}$ 18	$Z_{\bar{A}AN}$ 17
			N	Z_{ANM} 19	Z_{ANA} 21	$Z_{AN\bar{A}}$ 31	Z_{ANN} 145
	N	GENERATOR 2	M	Z_{NMM} 2	Z_{NMA} 3		Z_{NMN} 12
			A	Z_{NAM} 13	Z_{NAA} 17	$Z_{NA\bar{A}}$ 11	Z_{NAN} 89
			N	Z_{NNM} 55	Z_{NNA} 115		Z_{NNN} 1510

Table 3

Examples of Minimal Sufficient Statistics (MSS) Corresponding to the Parameters of the Loglinear Models of Section 4.

Parameter		MSS
(i)	θ	Z_{+++}
(ii)	θ_1	$2Z_{M++} + Z_{A++}$
(iii)	ρ_{11}	Z_{M++}
(iv)	ρ_{12}	$2Z_{MM+} + Z_{MA+} + Z_{AM+} + Z_{A\bar{A}+}$
(v)	γ_{12}	$2Z_{MM+} + Z_{MA+} + Z_{AM+} + Z_{AA+}$
(vi)	$(\rho\gamma)_{112}$	$2Z_{MM+} + Z_{MA+}$
(vii)	$(\rho\gamma)_{1122}$	Z_{MM+}

Table 4

Loglinear Models for Z-Table and MSS's of Corresponding Loglinear Models for the Y-Table

Parameters in Model for the Z-Table	Fitted Margins for Corresponding Loglinear Model for Y-Table*	Degrees of Freedom
(i) θ	Y_{++++++}	35
(ii) plus θ_1, θ_2, θ_3	[1] [1'] [2] [2'] [3] [3']	32
(iii) plus ρ_{11}, ρ_{22}, ρ_{33}	[11'] [22'] [33']	29
(iv) plus ρ_{12}, ρ_{13}, ρ_{23}	[11'] [22'] [33'] [12'] [1'2] [13'] [1'3] [23'] [2'3]	26
(v) plus γ_{12}, γ_{13}, γ_{23}	all two-way marginal tables	23
(vi) plus $(\rho\gamma)_{112}$, $(\rho\gamma)_{113}$, $(\rho\gamma)_{221}$, $(\rho\gamma)_{223}$, $(\rho\gamma)_{331}$, $(\rho\gamma)_{332}$	[11'2] [11'2'] [11'3] [11'3'] [122'] [1'22'] [22'3] [22'3'] [133'] [1'33'] [233'] [2'33']	17
(vii) plus $(\rho\gamma)_{1122}$, $(\rho\gamma)_{1133}$, $(\rho\gamma)_{2233}$	[11'22'] [11'33'] [22'33']	14

*All of the models also include a fixed 6-factor interaction corresponding to the positions of the doubled counts in the table.

Table 5

Various Loglinear Models Fitted to Data in Table 1

Model		df	G^2
(ii)	$\theta,\theta_1,\theta_2,\theta_3$	32	1599.0
(iii)	$\theta,\theta_1,\theta_2,\theta_3,\rho_{11},\rho_{12},\rho_{13}$	29	788.5
(iv)	$\theta,\theta_1,\theta_2,\theta_3,\rho_{11},\rho_{22},\rho_{33},\rho_{12},\rho_{13},\rho_{23}$	26	125.2
(v)	$\theta,\theta_1,\theta_2,\theta_3,\rho_{11},\rho_{22},\rho_{33},\rho_{12},\rho_{13},\rho_{23},$ $\gamma_{12},\gamma_{13},\gamma_{23}$	23	73.6
(vi)	$(\rho\gamma)_{112},(\rho\gamma)_{113},(\rho\gamma)_{221},(\rho\gamma)_{223},$ $(\rho\gamma)_{331},(\rho\gamma)_{332}$, plus all implied lower-order forms	17	35.13
(vii)	$(\rho\gamma)_{1122},(\rho\gamma)_{1133},(\rho\gamma)_{2233}$, plus all implied lower-order terms	14	24.3

Table 6

Parameter Estimates for Model (vii)
Fitted to the Galaskiewicz and Marsden Data from Table 1

Parameter	Estimate	
$\hat{\theta}$	-1.25	Grand Mean
$\hat{\theta}_1$	-2.34	
$\hat{\theta}_2$	-2.80	Choice
$\hat{\theta}_3$	-2.56	
$\hat{\rho}_{11}$	2.49	
$\hat{\rho}_{22}$	0.73	Symmetry
$\hat{\rho}_{33}$	1.82	
$\hat{\rho}_{12}$	0.90	
$\hat{\rho}_{13}$	0.78	Exchange
$\hat{\rho}_{23}$	0.94	
$\hat{\gamma}_{12}$	0.54	
$\hat{\gamma}_{13}$	1.05	Multiplex
$\hat{\gamma}_{23}$	0.06	
$(\hat{\rho\gamma})_{112}$	-0.07	
$(\hat{\rho\gamma})_{113}$	-0.14	Conditional
$(\hat{\rho\gamma})_{221}$	-0.10	
$(\hat{\rho\gamma})_{223}$	0.46	Asymmetry
$(\hat{\rho\gamma})_{331}$	-0.84	
$(\hat{\rho\gamma})_{332}$	0.32	
$(\hat{\rho\gamma})_{1122}$	-0.31	
$(\hat{\rho\gamma})_{1133}$	0.45	Multiplex Symmetry
$(\hat{\rho\gamma})_{2233}$	-2.23	

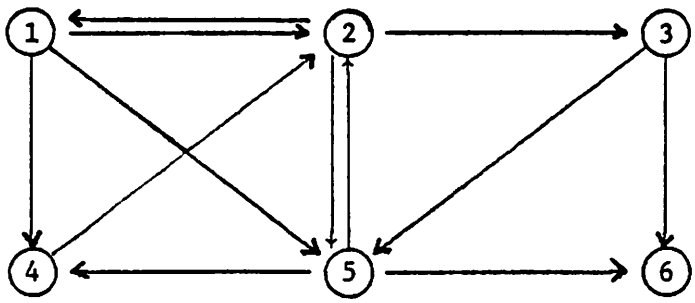

Figure 1. A univariate digraph with
g = 6 nodes and 12 directed arcs.

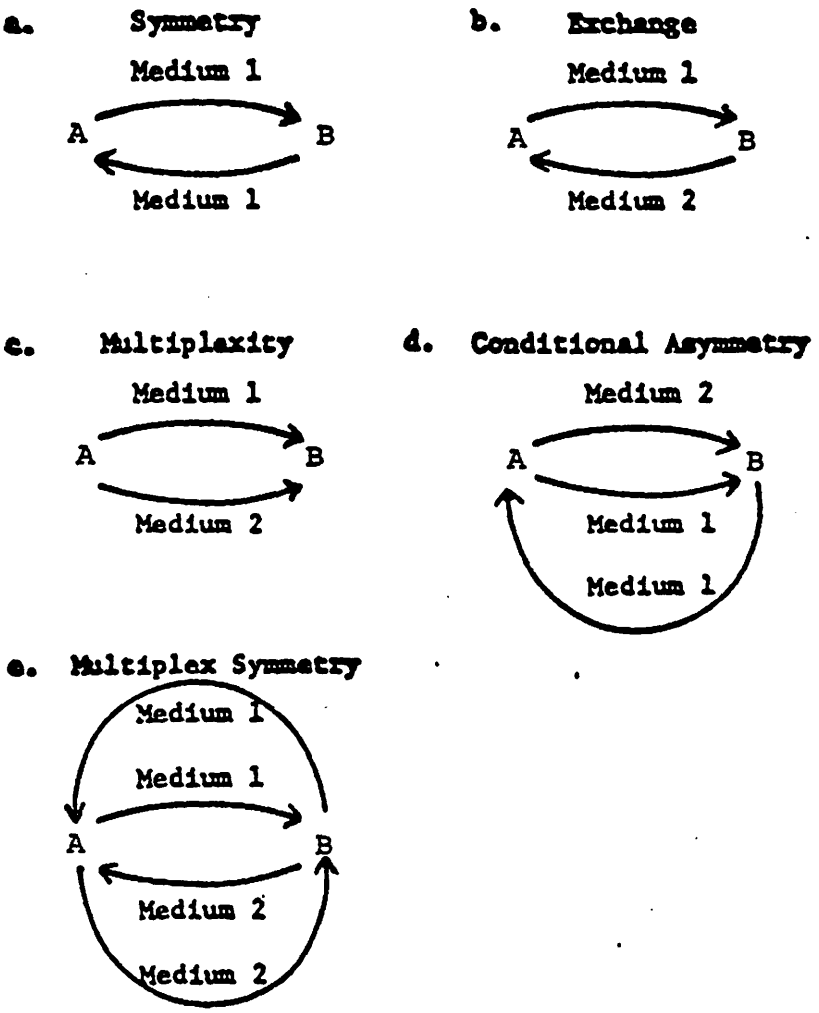

Figure 2. Patterns of flow dependency.
Source: Galaskiewicz and Marsden (1978)